UNITEX-PB, a set of flexible language resources for Brazilian Portuguese

Marcelo C.M. Muniz, Maria das Graças V. Nunes, Eric Laporte

UNITEX-PB, a set of flexible language resources for Brazilian Portuguese*

Marcelo C. M. Muniz[1] **, Maria das Graças V. Nunes**[1] **, Eric Laporte**[2]

[1]Núcleo Interinstitucional de Lingüistica Computacional – NILC
ICMC/USP Caixa Postal 668 – 13560-970 São Carlos, SP

[2]Institut d'électronique et d'informatique Gaspard-Monge – IGM
University of Marne-la-Vallée/CNRS - 5, bd Descartes – 77454 Marne-la-Vallée – France

marcelo@marcelomuniz.com.br, gracan@icmc.usp.br, Eric.Laporte@univ-mlv.fr

Abstract. *This work documents the project and development of various computational linguistic resources that support the Brazilian Portuguese language according to the formal methodology used by the corpus processing system called UNITEX. The delivered resources include computational lexicons, libraries to access compressed lexicons, and additional tools to validate those resources.*

1. Introduction

One of the main obstacles to the advancement of research, and consequently to the development of systems, in the field of natural language processing (NLP) in Brazil, has been the lack of language resources that, in the last analysis, provide all the domain-specific knowledge required in this field. Such resources are specialized, large and complex. Their construction requires trained interdisciplinary teams, and the cost of such work has prevented research in NLP on Portuguese to reach the same level as on certain other widely spoken languages.

Language resources required for developing applications provide linguistic knowledge. Some of them are compatible with various types of computer programs, like electronic lexicon, corpora and thesauri. Others are coupled with a specific tool such as a tagger or a morphological, syntactic or semantic analyzer. In both cases, the largest part of the time and effort required to develop a NLP application is dedicated to the construction of the language resources that support the operation of the application.

In the recent years, NLP researchers focused a lot of their effort on the standardization of the construction of such resources. Some standards and tools emerged at the international level and are being used by various research groups. One of these standards was developed at LADL (Laboratoire d'informatique documentaire et linguistique, University of Paris 7, France), the DELA (Dictionnaires électroniques du LADL), jointly with the corpus-processing tool INTEX [Silberztein, 2000a]. The DELA became the standard of electronic lexicon used by the informal research network Relex. These lexicons were used with INTEX, and now with its open-source counterpart, UNITEX [Paumier, 2002].

In this context, this article presents the construction of UNITEX-PB, a set of language resources for Brazilian Portuguese using the DELA standard and the UNITEX tool. In the next section, we introduce electronic lexicons and their modes of representation; in section 3 we describe the data formats of UNITEX-compatible lexicons; section 4 reports the design, development and validation of the language resources. Section 5 describes our

*Supported by CAPES, CNPq and FAPESP

contribution to the UNITEX software and section 6 contains final considerations on this work.

2. Lexicons for natural language processing

The notion of electronic lexicon is fundamental for most NLP systems and applications. An electronic lexicon is a data structure containing lexical items of a language and information about these items. Items can be isolated words such as *lua* "moon" and *mel* "honey", or sequences of words with an uncompositional meaning, e.g., *lua de mel* "honeymoon", *Casa de Cultura* "cultural centre" or *a grosso modo* "roughly". Information attached to lexical items includes values of morpho-syntactic features like part of speech, gender, number, degree, person, tense, mood, in addition to syntactico-semantic features like verbal or nominal sub categorization frames, and semantic links. However, definitions or associations with contextual representations are seldomly provided.

In the next paragraphs we recall basic notions and distinctions about NLP lexicons.

2.1. The MRD/MTD distinction

An electronic lexicon or machine-tractable dictionary (MTD), which can be used in NLP systems, is fundamentally different from a conventional dictionary, which is designed for human use and made by lexicographers. When a conventional dictionary is initially produced in digital format, or created in paper form but later transferred to a digital format, it is sometimes called a machine-readable dictionary (MRD). Lexical units are basically described in separate articles, with the classical structure in three parts: entry, category, definition. Numerous tentatives to derive MTDs from MRDs are recorded in literature [Wilks et al., 1988], but that trend did not produce any viable technology for the construction of electronic lexicon [Ide and Véronis, 1993]. Most large-coverage NLP lexicon of quality were produced from scratch: WordNet [Miller, 1985], DELA [Courtois, 1990] , NILC´s lexicon [Nunes, 1996], CISLEX [Maier-Meyer, 1995], DIGRAM [Ranchhod et al., 1999]. All these lexicons underwent a more or less smooth transition from construction to maintenance and are still in use today.

2.2. Lexical databases

A lexical database (LDB) is a computational structure designed to be able to support various types of knowledge on lexical units, and in particular links between distinct lexical units or between elements of distinct lexical units. One of the main features of LDBs, from the theoretical viewpoint, is that the lexicon is considered as a network of relations at various levels (morpho-syntactic, syntactic, semantic and paradigmatic). In other types of electronic lexicon, interlexical relations are not systematically nor exhaustively represented [Calzolari, 1990]. The representation and management of links in an LDB requires the use of a database management system, but lexical data are much more complex than most types of data taken into account in the field of databases [Ide and Véronis, 1994].

Examples of LDBs are WordNet [Miller, 1985] for English, and Diadorim, constructed for Brazilian Portuguese by the NILC team [Greghi et al., 2002], by merging information from the NILC's lexicon with information from a thesaurus of Portuguese [Dias-da-Silva et al., 2000]. The main purpose of Diadorim was to centralize all of this information into a unique database. Presently, it contains about 1.5 million lexical entries (simple words), represented in a relational database.

2.3. Libraries of language resources

At the end of the 80s, the reuse of existing lexicon and lexical databases became a hot topic, since reusing provides an obvious means of alleviating the initial effort for the

development of new applications [Evans and Kilgarriff, 1995]. The construction of standard exchange data formats, which began at that time [Normier and Nossin, 1990] and is still going on today [Lieske et al., 2001, Ide and Romary, 2002], certainly improved the reusability of lexicon.

However, reusing a lexicon implies it should have a certain degree of flexibility. A lexicon is not a static resource, it evolves with time. Due to the evolution of language across time, and especially of technical language, regular updates are necessary; a new application of a lexicon may involve the selection of a domain-specific vocabulary. Lexicon reuse is likely to be facilitated if it is implemented in the framework of a library of language resources (LRL). Whereas a language resource is a mere dataset, a library of language resources is a dynamic concept which also includes tools and data for maintenance and adaptation of the resources, and more generally for language resource management. Examples of LRLs are XELDA, INTEX [Silberztein, 2000b, Silberztein, 2000a] and UNITEX [Paumier, 2002]. INTEX and UNITEX support the DELA standard of lexicon formats and apply technologies designed at LADL. Most other general-purpose systems for language resource processing and text processing are deprived of even basic functionality for lexicon management. The design of the GATE system [Cunningham, 2002], for example, ensures that the user never performs any operation of language resource management, since every resource used has to be wrapped in a text-processing tool (ibid., p. 228). The open-source, general-purpose computational-linguistics toolkit NLTK [Loper and Bird, 2002, Bird and Loper, 2004] does not offer functionality for resource management either.

2.4. Automated inflection

In inflected languages, such as most European languages, the most basic notion in lexicon management is the distinction between lexicons of lemmas and lexicons of forms. Construction, maintenance and other lexicon management operations can be performed on a lexicon of lemmas, but not directly on a lexicon of forms, which contains much more redundancy and is too large to be conveniently edited. In all LDBs and emerging standards for lexicon formats, words are basically represented by lemmas. On the other hand, a lexicon of forms is adapted to text processing, since forms occurring in texts are directly described in lexical entries and these descriptions are accessible through efficient lookup. A lexicon of lemmas can also be used in applications, but this requires lemmatizing the text by applying a tagger or a stemmer, which gives only approximate results, or accessing the lexicon through morphological analysis, which is slower than direct lookup in a lexicon of forms.

The only way to combine the three constraints (flexible language resources, accurate results and computational efficiency) is to manage jointly a lexicon of lemmas and a lexicon of forms. The classical solution for this is to compile the first into the latter by automated inflection, i.e. generation of inflected forms from lemmas (e.g. [Domenig, 1988]; [Courtois, 1990]).

As far as emerging standards of lexicon formats are concerned, the inflected form/lemma duality has a place in OLIF data structures [Lieske et al., 2001], but the organization of data structures for inflected forms and inflection is completely left to users. The issue is scheduled in the ISO group on Language resource management [Ide and Romary, 2002], but has not been dealt with in priority.

In this context, few language engineering companies are aware of modern lexical resource management, and the NLP research community pays insufficient attention to issues such as the improvement of current techniques and their extension to new languages. The availability of more LRLs of quality would help designers and implementers

of applications to achieve a better integration between language resource management and their other activities. Thus, the advancement of technologies connected to LRLs, and the construction of resources for LRLs, are factors of progress in NLP.

3. UNITEX

UNITEX is an environment for linguistic resource development that can be used to parse texts of several million words in real time. The descriptions of the linguistic knowledge are formalized as electronic dictionaries, grammars represented as finite state graphs and lexicon-grammars.

Developers from different countries like France, Germany, Greece, Italy, Korea, Norway, Spain, Poland, Portugal and Thailand have also been working to build their own lexical dictionaries for the UNITEX/INTEX system[1].

UNITEX uses the formats and standards defined by the DELA system which was used as a source for the main international projects of definition of standards for electronic lexicon, from GENELEX [Normier and Nossin, 1990] to the ISO group on language resource management [Ide and Romary, 2002]. The ISO standard in construction has been implemented in the form of XML formats which are self-understandable and conform to other emerging standards. Software for converting the DELA format to/from these XML formats has been developed in the framework of the Outilex project. The DELA format contains the same information as the XML formats, and is less verbose, because it uses compact codes instead of self-understandable tags. In this paper, we give examples in the DELA format. This format allows to declare simple and compound lexical entries, which can be associated to grammatical information and inflection rules. These dictionaries are linguistic resources specifically designed to be used in automatic text processing operations. Variations of DELA include DELAF, which comprises inflected simple words, DELAC and DELACF, for non inflected and inflected compound words, respectively. DELAF and DELACF are automatically generated from DELAS and DELAC dictionaries.

The dictionaries of simple words (DELAS and DELAF) are simple lists of words associated to grammatical and inflectional information. The grammatical information is mainly of morphological type and corresponds to gender, number, degree, case, mood, tense, and person. However, the format makes it possible to gradually add syntactic and semantic information.

The lexical entries of DELAS have the following general structure:
(word), (formal description)
where *word* represents the *canonical form* (the lemma) of a simple lexical unit (in general, masculine singular for nouns and adjectives, and infinitive for verbs), and *formal description* corresponds to an alphanumeric code representing the attributes of an entry: its part of speech (optionally, subcategory), and its morphological behavior [Ranchhod, 2001]. The following entries are real examples of the Brazilian Portuguese DELAS dictionary:

mato, N001D026A01
beijar,V005
melodicamente,ADV

where *N* stands for noun, *V* for verb and *ADV* for adverb. The numerical code corresponds to the inflection rule. *Mato*, for example, is a noun, and its inflection rule for gender and number is 001. Codes *D* and *A*, followed by numerical codes, are used to allow the generation of the appropriate diminutive and augmentative, respectively. In

[1] See `http://www-igm.univ-mlv.fr/~unitex/`

case a word is of more than one grammatical class, there will be one DELAS entry for each class. For example, the inflected word *mato* should be generated by both DELAS entries: *matar,V030* and *mato,N001D026A01*.

The inflectional rules are formalized as a Finite State Transducer (FST) which associates sets of suffixes to the lexical DELAS entries and generates the corresponding inflected forms. For the entry *mato,N001D026A01*, the following inflected forms (DELAF entries) are produced:

matinho,mato.N:Dms
matinhos,mato.N:Dmp
mato,mato.N:ms
matos,mato.N:mp
matão,mato.N:Ams
matões,mato.N:Amp

The lexicon of compound words (here DELAC) constitutes a large part of the lexicon of any language. Compound words are sequences of words whose meaning cannot be derived from the meaning of their constituents. DELAC entries are similar to DELAS entries. The inflection rules represent restrictions on gender and number which cannot be accounted for by the morphological properties of their constituents. Thus, given the following entries of DELAC:

bom gosto,N+AN:ms - - (good taste)
ser(N010) humano(A001),N+NA:ms - +(human being)

the first compound noun, *bom gosto* has an internal structure *Adjective Noun* (AN); each entry is characterized by the possibility (+) or impossibility (-) of gender and number inflection, respectively. The elements of the compound that can be inflected receive the inflectional code that they have in the DELAS: N010 and A001 are the codes of number inflection rules, respectively, for both constituents of *ser humano*. The corresponding inflected compound words constitute the dictionary DELACF.

4. The construction of UNITEX PB

The process of developing the UNITEX system for Brazilian Portuguese was implemented in three steps: the design and implementation of the DELAS and DELAF dictionaries, the design and implementation of the DELACF dictionary, and the development of a library to access and manipulate the UNITEX-PB lexicon.

4.1. Design and implementation of Brazilian Portuguese DELAS and DELAF dictionaries

The starting point for the DELAS dictionary was the NILC‘s machine tractable lexicon which has over 1.500.000 entries and supports many NLP applications for Brazilian Portuguese, such as the Brazilian version of the grammar checker from MSOffice [Martins et al., 1998]. The text version of this lexicon was used as source of information in this phase.

In the one hand, since NILC´s lexicon contains all inflected words, we could automatically convert them to the DELAF format. A conversion process was proposed to achieve this goal. This filter deals with 14 grammar classes (noun, adjective, determiner, preposition, conjunction, numeral, pronoun, proper noun, verb, adverb, prefix, abbreviation, acronym, and interjection), which are the same set used by NILC´s lexicon. This process generated a lexicon with 1.542.563 inflected entries, wich we call *intermediate DELAF*. From that, the compound words and some verb inflections (*ênclises* and

mesóclises) from NILC´s lexicon had to be removed, since DELAF was not expected to contain them. Therefore, the number of entries decreased to 454.304.

In the other hand, the DELAF dictionary was expected to be generated from the DELAS dictionary, which should contain only lemmas and inflection rules. Thus, one had to pave the way for this by designing the inflection rules for the lemmas of NILC´s lexicon. A new version of the DELAF dictionary would be generated from them and it would be compared with *intermediate DELAF*.

In order to design the inflection rules, *intermediate DELAF* was divided into files, each one containing entries from only one grammar class. Some classes like adverb, conjuntion and interjection do not require inflection rules. A big challenge was to create the inflection rules for the other grammar classes that represent 99.99% of all entries in the dictionary.

For nouns and adjectives, the file of lemmas was divided into other files according to each lemma termination, for example, a file containing noun lemmas ending in ”o”, other containing noun lemmas ending in ”a”, and so on. This strategy aimed at simplifying the task by breaking it in small ones. For each file, an inflection rule was proposed. For example, for nouns ending in ”o”, the rule of Figure 1 was proposed. It is correct for most nouns of this category. This rule has two paths: in one, nothing is added to the lemma, and the associated inflectional code is *:ms*; in the other, the suffix *s* is added, and the associated inflectional code is *:mp*. These rules were manually generated and then automatically associated to all lemmas of each file. In a final step, all files were merged, originating the DELAS dictionary.

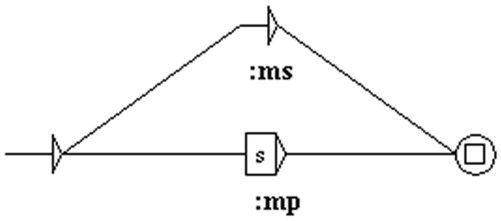

Figure 1: Number inflection rule for words ending in ”o”

In the next step, all the inflection rules defined in DELAS were automatically applied to the lemmas and the resulting file was compared to the file obtained from NILC´s lexicon (*intermediate DELAF*). A resulting file of errors was generated, i.e. a file containing invalid words produced by some inflection rule. These rules were manually analyzed and new versions of rules were associated to the lemmas in the DELAS dictionary. This process was repeated until no error was detected. It is interesting to note that this process allowed us to detect over a thousand mistakes in NILC´s lexicon, such as wrong inflections, omissions, spelling problems in lemmas, etc.

For the DELAS verbs, instead of using the verb set from NILC´s lexicon (containing 6.672 verbs), we adopted the list of 14.284 verbs and corresponding inflection rules, in a proper formalism, compiled by Vale [Vale, 1989]. This information was converted to the DELAS format by a conversion program. Finally, the inflected verbs of the DELAF were produced by the application of these inflection rules, completing the creation of the DELAF dictionary.

Table 1 shows the numbers of lemmas and inflection rules of the DELAS dictionary for Brazilian Portuguese, and Table 2 shows the final numbers of entries of the dictionaries of simple words, DELAS and DELAF.

Table 1: Number of Lemmas and Inflection Rules in DELAS.

Grammar Class	Number of inflection rules	Number of lemmas
Noun	378	32.178
Adjective	242	17.658
Determiner and Number	14	97
Preposition	17	95
Pronoun	35	67
Verb	102	14.284
Others	5	804

The final DELAF dictionary was then compacted using UNITEX and all the resources of the system could be used for processing texts in Portuguese.

Table 2: Statistics about DELAS and DELAF.

Dictionary of Simple Words - DELAS		
Total number of entries		67.466
Dictionary of Inflected Words - DELAF		
Class	Entries	Lemmas
Abbreviation	214	214
Adjective	59.349	17.658
Adverb	2.628	2.628
Determiner	8	2
Conjunction	44	44
Interjection	23	23
Number	238	95
Prefix	55	55
Preposition	201	95
Pronoun	266	67
Acronym	468	468
Noun	71.285	32.178
Verb	743.316	14.284
Total	878.095	61.135
Dictionary size		48.8 MB
Dictionary size (compacted)		0.99 MB
Compression tax		97.9 %

By way of curiosity, the total number of inflected forms in the DELAF dictionary is over 93% higher than NILC´s lexicon, which contains 454.304 inflected simple forms. Considering that NILC´s dictionary has proved to be a valuable resource for practical applications, we can state that the UNITEX dictionaries for Brazilian Portuguese are able to support significant investigations over corpora.

4.2. Design and implementation of Brazilian Portuguese DELACF

Similarly to DELAS and DELAF, DELAC is the UNITEX dictionary which contains the lemmas of compound words associated to the respective inflection rules and DELACF is the result of the application of the inflection rules, i.e. the set of all inflected compound words.

By the time of the construction of these resources, the formalism for converting DELAC to DELACF was not totally formalized by the European developers of UNITEX. Therefore, we decided to directly build DELACF (for which the format is well known) from the inflected compound words from NILC´s lexicon, through the use of a conversion program.

An example of an entry in the DELACF dictionary is the following:
rabos-de-tatu,rabo-de-tatu.N+NDN:mp
where NDN stands for noun + preposition *(de)* + noun. As the information about the parts of speech of the constituents of the compounds was not available in NILC´s lexicon, they were manually inserted.

The final statistics about the DELACF dictionary are presented in Table 3.

Table 3: Statistics about the Dictionary of Compound Nouns - DELACF

DELACF	
Total number of entries	4.077
Total number of lemmas	2.009
Dictionary size	301 KB
Compacted dictionary size	121 KB

5. Contributions to the UNITEX software

In order to contribute to the UNITEX software, we have extended some functionalities of the UNITEX tool under the terms of the LGPL[2] licence.

A library of simple functions to access and manipulate the compressed lexicons was implemented and it is used independently of UNITEX. With this library, it is possible to load the dictionary in memory, perform a search for a word and unload the dictionay. The search function returns an empty string if it doesn‘t find the word, or returns all the information associated to that word. This library was developed in two programming languages: ANSI C and Java.

A program called *Dicionário* was implemented as an example of how to use the library. This is a simple program that can load any compressed dictionary in DELA format and has an interface to perform searches for words in the dictionary. This program together with the library to access compressed lexicons, as well as all the resources built in this project, can be downloaded from:

http://www.nilc.icmc.usp.br:8180/unitex-pb/

Some applications which use DELAF were implemented and one of them, a morpho-syntactic tagger, allows the user to tag a text with all information (tags) from the DELAF dictionary. The user can enter his text in a form or by uploading a file and also has the option to choose which tags s/he wants: canonical, grammatical, semantic attributes and morphologic attributes. The result can be displayed as a HTML page, a TXT file or a compact TXT file. This tagger is similar to the tool constructed by the researchers from LabEL, Portugal, called ANELL[3].

The integration of these software components into UNITEX is under study. This integration would allow UNITEX users interested in other languages to benefit from their implementation.

[2]See `http://www.gnu.org/licenses/lgpl.html`
[3]See `http://label10.ist.utl.pt/anell/`

6. Conclusion

Our main objective was the construction of language resources for Brazilian Portuguese in the UNITEX framework. The impact of this work is at several levels. The number of simple-word entries of the NILC's lexicon has increased by 93%. A complete set of inflection rules for Brazilian Portuguese simple words was created in the form of a set of transducers, viewable and editable in graphical form. Brazilian Portuguese was the first language for which such a resource is available.

Software components for access to compressed lexicon and presentation of lookup results were implemented. If they are integrated into UNITEX, they will be an a decisive contribution to lexicon management.

This work was interdisciplinary and involved a close co-operation between linguists and computer scientists. It also contributed to confirm NILC as a member of the RELEX network.

The resources have been made freely available in recent releases of UNITEX with the LGPL-LR license. This decision gives researchers, companies and other users a larger access to Brazilian Portuguese language resources, and more opportunity to apply text processing techniques.

In addition, the results validate our approach to language resource management. This approach can be extended to other existing inflected-form lexicons in other languages.

References

Bird, S. and Loper, E. (2004). NLTK: the natural language toolkit. *ACL 2004*. `http://www.ldc.upenn.edu/sb/home/papers/nltk.pdf`.

Calzolari, N. (1990). Structure and access in a automated lexicon and related issues. In *Linguistica Computazionale vol. II - Computational Lexicology and Lexicography: Special Issue dedicated to Bernard Quemada*, pages 139–161. Pisa, Giardini Editori e Stampatori.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français, *Langue Française. Dictionnaires électroniques du français*, 87:11–22.

Cunningham, H. (2002). Gate, a general architecture for text engineering. *Computers and the Humanities 36*, pages 223–254.

Dias-da-Silva, B. C., Oliveira, M. F., Moraes, H. R., Hasegawa, R., Amorim, D., Paschoalino, C., and Nascimento, A. C. (2000). A construção de um thesaurus eletrônico para o português do Brasil. In *V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000)*, pages 1–11. Atibaia, SP.

Domenig, M. (1988). Word manager: A system for the definition, access and maintenance of lexical databases. In *Proceedings of Colling'88*. Budapest.

Evans, R. and Kilgarriff, A. (1995). MRDs, standards and how to do lexical engineering. Technical Report ITRI-95-19, University of Brighton.

Greghi, J. G., Martins, R. T., and Nunes, M. G. V. (2002). Diadorim: a lexical database for Brazilian Portuguese. In *Proceedings of the Third International Conference on language Resources and Evaluation. LREC2002.*, volume IV, pages 1346–1350. Las Palmas, Ilhas Canárias.

Ide, N. and Romary, L. (2002). Standards for language resources. In *Proceedings of the Third International Conference on language Resources and Evaluation. LREC2002.*, volume IV, pages 839–844. Las Palmas, Ilhas Canárias.

Ide, N. and Véronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In *KB&KS Workshop*. Tokyo.

Ide, N. and Véronis, J. (1994). A feature-based data model for lexical databases. In *Hockey, S., Ide, N. Research in Humanities Computing IV*, pages 193–206. Oxford University Press.

Lieske, C., McCormick, S., and Thurmair, G. (2001). The open lexicon interchange format (OLIF) comes of age. *Machine Translation Summit VIII*. `http://www.eamt.org/summitVIII/papers/lieske.pdf`.

Loper, E. and Bird, S. (2002). NLTK: the natural language toolkit. In *ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia.

Maier-Meyer, P. (1995). Lexikon und automatische Lemmatisierung. Dissertation, Universität München.

Martins, T. B. F., Hasegawa, R., Nunes, M. G. V., , Montilha, G., and Oliveira Jr, O. N. (1998). Linguistic issues in the develepment of ReGra: a grammar checker for Brazilian Portuguese. *Natural Language Engineering*, 4(4):287–307.

Miller, G. A. (1985). Wordnet: a dictionary browser. In *Proceedings of the First International Conference on Information in Data*. University of Waterloo.

Normier, B. and Nossin, M. (1990). Genelex project: Eureka for linguistic engineering. In *Proceedings of the International Workshop on Electronic Dictionaries*, pages 63–70. OISA, Kanagawa, Japan.

Nunes, M. e. a. (1996). O processo de construção de um léxico para o Português do Brasil: Lições aprendidas e perspectivas. In *II Encontro para o Processamento Computacional de Português Escrito e Falado*, pages 61–70. CEFET-PR, Curitiba.

Paumier, S. (2002). *Manuel d'utilisation du logiciel Unitex*. IGM, Université de Marne-la-Vallée. `http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf`.

Ranchhod, E., Mota, C., and Baptista, J. (1999). A computational lexicon of Portuguese for automatic text parsing. In *Proceedings of SIGLEX99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*, pages 74–80. College Park, Maryland, USA.

Ranchhod, E. M. (2001). *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações*, chapter O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais, pages 13–47. Caminho, Lisboa.

Silberztein, M. (2000a). Intex: a FST toolbox. *Theoretical Computer Science*, 231:33–46.

Silberztein, M. (2000b). *INTEX Manual*. LADL, Université Paris 7.

Vale, O. A. (1989). *Constitution d'un dictionnaire électronique de conjugaison des verbes du portugais du Brésil - Rapport LADL n°27*. Université Paris 7.

Wilks, Y., Fass, D., Guo, C.-M., McDonald, J., Plate, T., and Slator, B. (1988). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceedings of Colling'88*, pages 750–755. Budapest.